Electronic Research Archive

http://www.aimspress.com/journal/era

ERA, 32(4): 2848–2864.
DOI: 10.3934/era.2024129
Received: 17 October 2023
Revised: 19 January 2024
Accepted: 29 January 2024
Published: 11 April 2024

Research article

Skeleton action recognition via graph convolutional network with self-attention module

Min Li[1], **Ke Chen**[1], **Yunqing Bai**[2,*] **and Jihong Pei**[2]

[1] School of Mathematical Sciences, Shenzhen University, Shenzhen 518060, China

[2] ATR National Key Laboratory of Defense Technology, Shenzhen University, Shenzhen 518060, China

* **Correspondence:** Email: 1800201011@email.szu.edu.cn; Tel: +8613428919322; Fax: +86075526530939.

Abstract: Skeleton-based action recognition is an important but challenging task in the study of video understanding and human-computer interaction. However, existing methods suffer from two deficiencies. On the one hand, most methods usually involve manually designed convolution kernel which cannot capture spatial-temporal joint dependencies of complex regions. On the other hand, some methods just use the self-attention mechanism, ignoring its theoretical explanation. In this paper, we proposed a unified spatio-temporal graph convolutional network with a self-attention mechanism (SA-GCN) for low-quality motion video data with fixed viewing angle. SA-GCN can extract features efficiently by learning weights between joint points of different scales. Specifically, the proposed self-attention mechanism is end-to-end with mapping strategy for different nodes, which not only characterizes the multi-scale dependencies of joints, but also integrates the structural features of the graph and an ability of self-learning fusion features. Moreover, the attention mechanism proposed in this paper can be theoretically explained by GCN to some extent, which is usually not considered in most existing models. Extensive experiments on two widely used datasets, NTU-60 RGB+D and NTU-120 RGB+D, demonstrated that SA-GCN significantly outperforms a series of existing mainstream approaches in terms of accuracy.

Keywords: skeleton-based action recognition; self-attention module; graph convolutional network

1. Introduction

Human action recognition is one of the main research areas in the study of computer vision and machine learning, aiming to understand and assign a label to each action [1, 2]. It appears in a wide range of applications such as video surveillance systems [3, 4], human-computer interaction and robotics for

human behavior characterization [5].

In general, most existing approaches in human action recognition fall into two main categories: RGB video based approaches [6–8] and skeleton-based methods [9–15] with respect to the types of input data. The RGB video based approaches are computationally expensive in processing RGB pixel information of videos, while the skeleton-based models are more efficient as they only need to calculate the 2D or 3D human joint point data. In addition, skeleton-based action recognition stands out due to its great potential in adaptability to dynamic circumstance and illumination variation [16–20].

It is well-known that extracting effective features for background separation and representation learning in skeleton-based action recognition is a challenging problem. The traditional deep learning-based methods are good at assembling skeletons as a set of independent joint features [21, 22] or pseudo-images [23, 24], i.e., representing skeleton data as vector sequences or two-dimensional vectors. With the constructed features, models and spatial-temporal correlations between them are given and learned. However, such vector-based representation cannot fully exploit the intrinsic relationships of human joints on account of ignoring the dependencies and inherent connections of the related ones. Thereupon, graph convolution networks (GCN) [25] have been used to capture inter-joint interactions, in which the natural connections of human joint points can be modeled as a spatial graph that corresponds to a graph convolution layer. As such, GCN-based spatial-temporal model is built at different time steps.

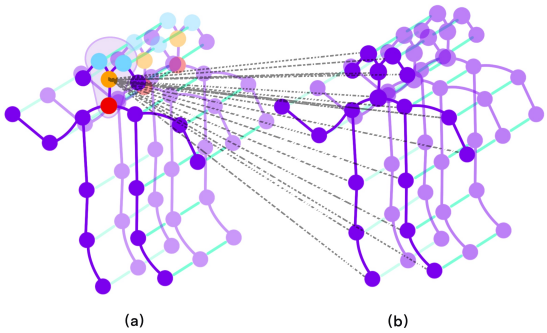

Figure 1. (a) Illustration of spatial-temporal GCN. (b) Illustration of mapping strategy used in SA, which exhibits the possible positive connections between different nodes.

For robust action recognition, the GCN-based skeleton graph prefers to extract multi-scale structural features, which are set subjectively [25]. However, the artificially designed graph structures involve connections that may hinder human actions in actual motion, resulting in models that are insufficient to effectively capture the spatio-temporal joint dependencies of complex regions. For example, there is usually a close relationship between hands and legs in the "running" action. However, it is difficult to obtain this relationship by an artificially constructed multi-scale structure. The recent approaches based on the self-attention mechanism (SA) can be used to learn the interactions between joints [26,27]. However, these methods only verify the feasibility of SA experimentally. To address these problems, we propose a unified framework, namely SA-GCN, that integrates the SA into the GCN for low-quality motion video data with fixed viewing angle*. Compared with existing methods, this model

*This dataset includes the raw data without regularization, which is introduced in ST-GCN [25].

fully exhibits the possible positive connections between different nodes, as shown in Figure 1.

To the best of our knowledge, this is the first study which pursues the impact of the SA mechanism on GCN in the area of action recognition (the latest and most representative study on SA-augmented GCNs was mainly to tackle the over-fitting and over-smoothing problems, e.g., see [28]). On the one hand, SA-GCN uses GCN as the basic component of the attention mechanism to extract the weights of joint point data in different frames of the same video. On the other hand, this component is beneficial to deepen the correlations of disconnected nodes of human joints, so that the learned long-distance temporal relationship is reflected in distant frames.

Extensive experiments on NTU-RGBD [29] and NTU120 [30] datasets demonstrate a superior performance of SA-GCN in terms of accruacy, and in comparison with a range of state of the art models across various skeleton-based action recognition tasks.

In summary, this paper claims three main contributions:

1) SA-GCN is a unified spatio-temporal graph convolution framework, which extracts better features by automatically learning the weight information between joint points at different scales. In addition, SA-GCN introduces a new SA mechanism with different scales for characterizing the multi-scale dependencies of joints, which greatly promotes the effectiveness in the learning of spatio-temporal features.

2) The proposed SA mechanism combines the structural informtion of the graph and the self-learning fusion features.

3) The proposed model achieves competitive performance for skeleton-based action recognition.

2. Related work

2.1. Graph-based neural networks

Graph neural networks (GNNs) have shown a great potential in computing irregular nodes in a graph [31]. Existing GNN models are mainly divided into two categories: spectral-based methods [32] and spatial-based approaches [27]. The former is based on feature decomposition, and performs convolution operations in the graph spectral domain. Thus the computational efficiency is limited due to the large amount of computation brought by feature decomposition. The latter (i.e., spatial-based GNNs) directly implements the operation of convolution on graph nodes and their neighbors in the spatial domain, thus is able to overcome the limitations of spectral-based methods. The method presented in this paper belongs to the second category.

Multi-scale GNNs are often used to capture the features for adjacent points. For example, [33] obtains image features by aggregating the higher-order polynomials of graph adjacency matrices. [34] represents multi-scale features by utilizing the graph adjacency matrix with a higher power. Nonetheless, these approaches inevitably and undesirably involve the artificial setting of structures for the adjacent points. Moreover, it is insufficient to rely only on the experimental verification that the setting of structures is optimal. Hence, the SA mechanism is proposed to lighten the problem of human interference.

2.2. SA mechanism

Attention mechanism [26] is critical in human cognition, which has revolutionized natural language processing and has been applied in image recognition [35], image synthesis [36] and video processing [37]. It is well-known that the human visual system captures structural features through partial glimpse and salient objects, rather than processing the whole at once [38].

Recently, attention has been proposed to improve the performance of convolutional neural networks (CNNs) for large-scale classification tasks. For example, a residual attention network (RAN) was introduced in an end-to-end training fashion [39], in which the attention modules are stacked and attention-aware features change adaptively as layers go deeper. Similarly, the channel-wise attention is incorporated into squeeze-and-excitation networks, which adaptively recalibrates the global average-pooled features by explicitly using interdependencies between channels [40].

However, the attention mechanisms have rarely been used in GCNs for action recognition. Considering the proven advantages of attention in computer vision tasks [41], it is natural to integrate SA into GCN as done in this paper, which reinterprets mathematically the positional relationship of each feature fusion in the attention mechanism. Experiments of large-scale dataset further demonstrate that SA mechanism can bring improvements in performance.

2.3. Skeleton-based action recognition

Traditional methods for skeleton-based action recognition can be divided into two types: handcrafted feature-based models [42] and deep learning-based approaches [43]. However, the performance of handcrafted feature-based methods is often unsatisfactory since local factors are used and the most important structural connections of the human body are ignored. The emergence of deep learning-based methods significantly improve the performance of skeleton recognition. For example, spatial graph convolution and interleaved temporal convolution are used for spatio-temporal characterization of the skeleton data (ST-GCN) [25]. Similarly, a multi-scale module is introduced to implement graph convolution modeling by raising the graph adjacency matrix of skeleton to a higher power [44]. In addition, the model of multi-scale adjacency matrix can also be performed by generating human poses and adding spatial graph convolution [45]. In [27], by using video frame segmentation, a set of videos is divided into four groups according to time and space for processing and combination, which expands the receptive field and achieves a great success in computer vision.

Different from the above methods, the graph convolution method proposed in this paper uses a SA mechanism. More specifically, the proposed approach uses GCN as part of the SA mechanism instead of forcing them to be summed as in existing methods [26]. In this way, the proposed method can better re-express the weights of the connection matrix.

3. GCN with SA mechanism

In this section, we present the proposed SA-GCN framework in detail.

3.1. ST-GCN for skeleton data

Skeleton data usually consists of a sequence of frames, in which each frame has a set of human joint coordinates in 2D or 3D form [25]. Naturally, a spatial-temporal graph for skeleton data can be

constructed, in which the nodes are human joint points and the edges include the natural connections of joints in human body structure over time in motion.

The skeleton data is formulated as a undirected spatial temporal graph G=(V, E), where $V = \{v_{ti}|t = 1, ..., T, i = 1..., N\}$ is the set of nodes including all the joints, T denotes video frames which feature both intra-body and inter-frame connectons, N is the total number of joints and E is the edge set. Usually, E is split into two subsets. One represents the intra-skeleton connection in each frame and the other depicts the inter-frame edges, which corresponds to the same joint point connections in consecutive frames. In fact, the ST-GCN is composed of spatial graph convolution and temporal graph convolution.

In the spatial dimension, the graph convolution is manipulated on each node and the corresponding neighbor set. For each node v_{ti}, the neighbor set is given as:

$$B_S(v_{ti}) = \{v_{tj}|d(v_{tj}, v_{ti}) \leq D\},$$

where $d(v_{tj}, v_{ti})$ is the shortest distance of any path from v_{tj} to v_{ti}. $D = 1$ means the 1-neighbor set of joint nodes. That is, if there is a connection between v_{tj} and v_{ti}, then $d(v_{tj}, v_{ti})$ is 1, and if there is no connection, it is 0.

According to the spatial partition strategies introduced in [25], the above neighbor set is divided into three labels, including: 1) the root node itself; 2) the centripetal subset which contains the adjacent nodes closer to the center of gravity than the root; 3) the centrifugal subset, which includes the adjacent nodes that are further away from the center of gravity than the root, as shown in (a) of Figure 1 (red, blue and orange represent three division strategies respectively). Hence, the spatial graph convolution is defined as:

$$f_{out}(v_{ti}) = \sum_{v_{tj} \in B_S(v_{ti})} \frac{1}{Z_{ti}\left(v_{tj}\right)} f_{in}\left(v_{tj}\right) \cdot W\left(L_{ti}(v_{tj})\right) \tag{3.1}$$

where $f_{in}(\cdot)$ and $f_{out}(\cdot)$ represent input and output features of this convolution layer, v_{tj} denotes the *j-th* node of the graph on frame t, and $B_S(v_{ti})$ is the sampling area of the convolution, characterized by the vertex with a distance of 1 from the target vertex. $W(\cdot)$ is the weighting function similar to convolution, which gives a weight vector based on the given data. $L_{ti}(\cdot)$ is the label function, which allocates a label from 0 to $K - 1$ to each node in $B_S(v_{ti})$. Usually, we can set K=3 [25]. $Z_{ti}(\cdot)$ denotes the number of nodes in the subset of $B_S(\cdot)$ with the label L_{ti}.

In practice, the skeletal sequence is usually denoted as a 3-order tensor of shape $C \times T \times N$, where C, T and N are respectively the numbers of channels, frames and joints. Thus, the implementation of Eq (3.1) can be transformed into the tensor format, such as $f_{in} \in \mathbb{R}^{C_{in} \times T \times N}$ and $f_{out} \in \mathbb{R}^{C_{out} \times T \times N}$ that represent the input and output feature maps, respectively. To perform multiplication of tensors, we reorder the elements of a tensor into a matrix through unfolding or flattening.

More technically, the connections between nodes in the skeleton graph are represented as an adjacency matrix $A \in \{0, 1\}^{N \times N}$. The adjacency matrix corresponding to the *k-th* subset of the neighbor set $B_S(v_{ti})$ can be denoted as A_k.

Let $F_{in} \in \mathbb{R}^{N \times C_{in}}$ be the input features of all joints in each frame, where C_{in} is the dimensionality of the input feature, and $F_{out} \in \mathbb{R}^{N \times C_{out}}$ is the output feature from the spatial graph convolution, where C_{out} is the dimensionality of output feature. Therefore, the implementation of the above Eq (3.1) can be rewritten as:

$$F_{out} = \sum_{k=0}^{K-1} \mathrm{M}_k \odot \widetilde{A}_k F_{in} \mathrm{W}_k, \tag{3.2}$$

where $\widetilde{A}_k = \mathbf{\Lambda}_k^{-\frac{1}{2}}(\mathbf{A}_k + I)\mathbf{\Lambda}_k^{-\frac{1}{2}} \in \mathbb{R}^{N\times N}$ is the normalized adjacent matrix for each partition label. $\mathbf{\Lambda} \in \mathbb{R}^{N\times N}$ is the diagonal degree matrix with $\Lambda_k^{ii} = \sum_j \left(A_k^{ij}\right) + \alpha$, and α is a small number to avoid division by zero. $\odot$ is the element-wise multiplication. $\mathrm{M}_k \in \mathbb{R}^{N\times N}$ and $\mathrm{W}_k \in \mathbb{R}^{C_{in}\times C_{out}}$ are learnable weight matrices for each partition label, which capture the strength of connections (i.e., edge weights) and the feature importance respectively.

In the temporal dimension, since the graph is given by corresponding joints in the two adjacent frames, the number of neighbors for each node is 2. Accordingly, the temporal graph convolution is similar to the classical convolutional operation, which is performed on the $T \times N$ output feature matrix mentioned above. Generally, the kernel size is set to 9, as in [10, 11, 25]. However, it should be emphasized that the respective fields for the spatial and temporal dimensions are set artificially. Thus, it is necessary that a SA mechanism is enforced to reduce the dependance on human interference.

3.2. SA module

As illustrated in the previous Section 3.1, the graph convolution can be obtained by a standard 2D convolution and a multiplication between the resulting tensor and the normalized adjacency marix under the spatial temporal cases. To automatically capture the relationships within the matrix, we introduce a SA operation into ST-GCN to further focus on the important regions and increase the model's attention to important joint point information.

The SA function can be viewed as a mapping from the query and a set of key-value pairs to an output [8,26]. For an input tensor of shape (C_{in}, T, N), we flatten it to the corresponding matrix $X \in \mathbb{R}^{C_{in}T\times N}$ and perform the single head attention. Hence, the output of the SA mechanism can be formulated as:

$$Attention(Q, K, V) = softmax(\frac{(QK^T)}{\sqrt{d_k}})V, \tag{3.3}$$

where $W_Q, W_K \in \mathbb{R}^{C_{in}\times d_k}$ and $W_V \in \mathbb{R}^{C_{in}\times d_k}$ are learned linear transformations that map the input X to queries $Q = XW_Q$, keys $K = XW_K$ and values $V = XW_V$. d_k is the depth of queries and keys.

Specifically, SA explores the correlations in a skeletal sequence, which are calculated for each position by the weighted sum over all positions. Moreover, the weight of each position in the similarity matrix is dynamic. Let x denote the input signal and y be the output. The SA module can be written as:

$$y(x) = \sigma(h(f(x) \cdot g(x)) + x), \tag{3.4}$$

where $f(x) = softmax(\theta(x)^T\phi(x))$ is used to generate the similarity matrix. $\theta(x) = W_\theta x$, $\phi(x) = W_\phi x$, $g(x) = W_g x$, $h(x) = W_h x$, and W_h, W_θ, W_ϕ and W_g are learnable. $\sigma(\cdot)$ is the activation function. In fact, here the similarity matrix is a weighted adjacent matrix in Eq (3.2). Hence, Equation (3.4) is named as SA-GCN.

3.3. Spatial graph convolution of SA-GCN

It is common sense for each action recognition category that the motion and nodes in different frames are different. Therefore, it is not enough to express the intra-skeleton connections in each frame and the inter-frame correlations by only learning a total joint point weight matrix as in ST-GCN

(Eq (3.4)). To address this problem, we propose to learn and represent all weights of the node for each frame by the attention mechanism.

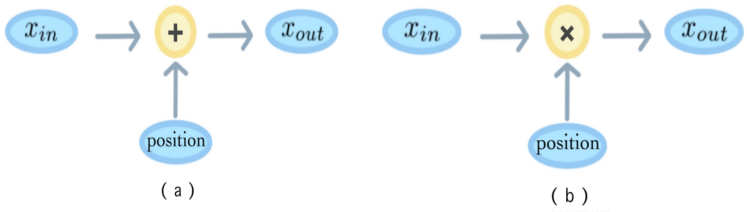

Figure 2. The difference for transform architecture between traditional methods and the proposed approach.

In traditional methods of attention, the input is usually expressed as the addition of the input matrix and the position variable, as shown in Figure 2(a). It is formally defined as:

$$x_{out} = x_{in} + A \tag{3.5}$$

where A can be represented as the positional relationship of the input nodes, and x_{in} and x_{out} represent the original data and the fusion data with the position variable.

However, in this paper we do not use the simple addition to fuse the input data and location variables. Instead, as shown in Figure 1(b), feature fusion is expressed as:

$$x_{out} = D_{norm} x_{in}, \tag{3.6}$$

where x_{in} and x_{out} represent the input and output matrices of the attention mechanism. D_{norm} denotes the Laplacian regularization matrix, which represents the position information that can be formulated as $D_{norm} = \mathbf{\Lambda}^{-\frac{1}{2}}(\mathbf{A} + \mathbf{I})\mathbf{\Lambda}^{-\frac{1}{2}}$. A is the adjacency matrix, I is the identity matrix, and $\mathbf{\Lambda}$ is the degree matrix of the node. In this paper, D_{norm} is implemented as $\mathrm{M}_k \odot \widetilde{A_k}$.

Compared to the traditional methods, the regularization matrix D_{norm} in Eq (3.6) guarantees the adaptive choice of the position matrix instead of manual setting. Simultaneously, we utilize the multiplication between input matrix and the position matrix (i.e., $D_{norm} f_{in}$), rahter than addition of the two in the traditional transformer architecture.

Next, we process the resulting matrix of Eq (3.6) to find the weight relationship between the different nodes of the input matrix. By integrating the attention module into Eq (3.6), the output result can be converted into multiple data matrices containing different information streams, as shown in Figure 3. Therefore, the conversion formula for input data is denoted as:

$$\begin{aligned} Q &= D_{norm} x_{in} W_Q, \\ K &= D_{norm} x_{in} W_K, \\ V &= D_{norm} x_{in} W_V, \end{aligned} \tag{3.7}$$

where W_Q, W_K and W_V are the corresponding parameter matrices. d_k in Eq (3.3) is defined as the row vector of the input data, that is, the number of joint points. Specifically, it is proved experimentally that using batchnorm in Eq (3.3) is more effective than the softmax function.

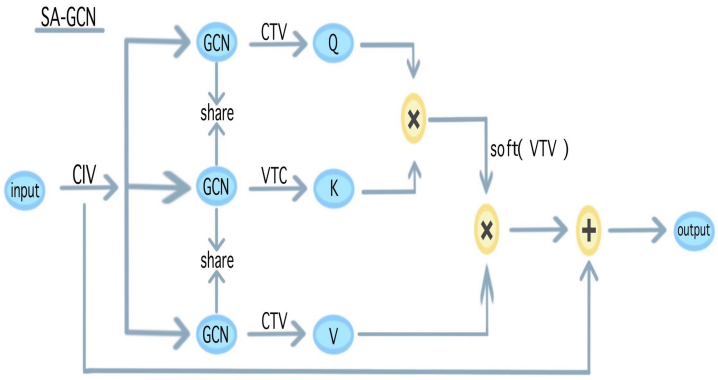

Figure 3. The flow chart of attention mechanism, which illustrates how input matrix is converted to a data matrix with different information streams.

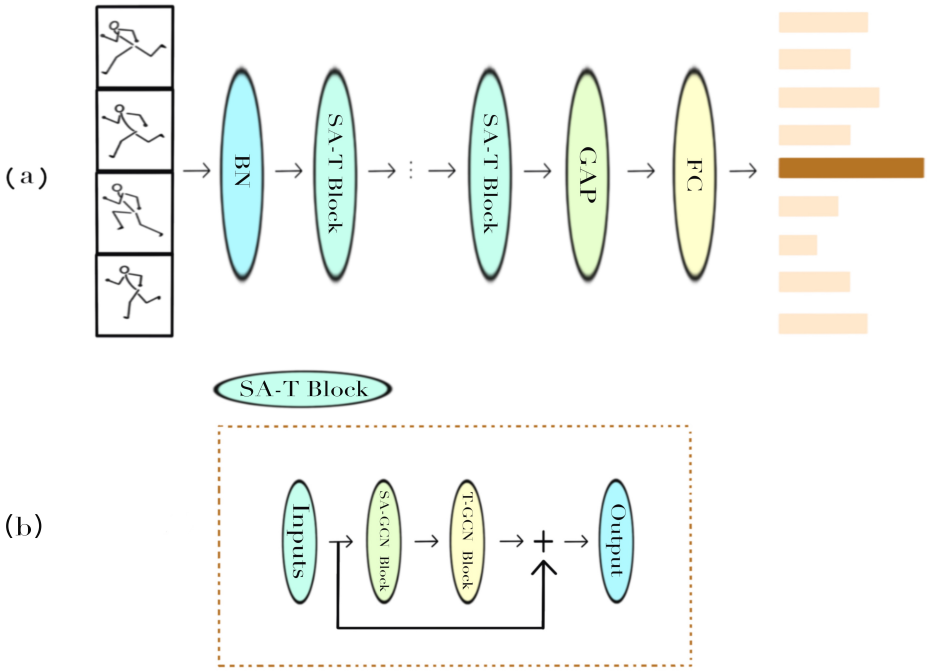

Figure 4. The spatio-temporal flow chart of SA-GCN.

We then set $W_{att} = Attention(Q, K, V)$. By using Eqs (3.3), (3.4) and (3.6), the proposed network can be written as:

$$f_{out} = \sigma(W_{att}). \tag{3.8}$$

We can see from Eq (3.8) that SA-GCN assembles a single module of the attention mechanism in a form similar to the spatial GCN (Eq (3.4)). Moreover, the proposed attention mechanism is heuristic since the mathematical formula of Eq (3.7) coincides with the input form Eq (3.6) proposed in this paper. The flow chart of SA-GCN is generalized in Figure 4, in which the proposed weight matrix of each frame is combinated with the original joint point weight matrix M_k of ST-GCN in Eq (3.2).

An important theoretical advantage of SA-GCN is that the attention mechanism in SA-GCN is interpretable. Specifically, the standard GCN is used as a component of the attention mechanism, which has a more theoretical derivation in the processing of data location variables. This enhances the dependency of the data on the existence of different joints. Specifically, the existing methods only describe the relationship between key nodes and adjacent joint points. However, the features processed by the proposed SA in this paper more fully demonstrate the positive influence between key points and all other nodes involved in the action (for example, the mapping strategy from (a) to (b) shown in Figure 1).

3.4. Temporal graph convolution of SA-GCN

The previous subsection (i.e., Section 3.3) is focused on improving the spatial method in SA-GCN, for which the attention mechanism is used to extract the joint weights of different frames in a video. On the other hand, it is well recognized that temporal modeling is also essential for video action recognition. For temporal modeling, we adopt the classical 2D graph convolution with kernel size 9 as introduced in the original ST-GCN [25]. That is

$$f_{out} = Cov2D[K + 1][f_{in}]. \tag{3.9}$$

where $D = 1$ and $K = 9$.

3.5. Spatial temporal graph convolution of SA-GCN

To deploy the proposed SA-GCN on video data, we build a concrete network structure upon the spatio-temporal graph convolution, as shown in Figure 4(a). Specifically, the network utilizes 10 SA-GCNs to perform feature fusion and feature extraction on the input spatio-temporal skeleton data. And the feature information is aggregated by global average pooling. Finally, the prediction result is given by the fully connected layer and the softmax layer. Among them, each SA-T block in (b) of Figure 4 is composed of a temporal graph convolution and a spatial graph convolution respectively, to extract a mixture of spatio-temporal features, and it also uses the residual structure to to prevent data overfitting and improve the generalization ability of the model, as shown in Figure 4(b).

In summary, SA-GCN jointly learns the total weight matrix and the weight matrix of each frame, which are expected to greatly improve the performance of the original ST-GCN. The next section will present the experimental results and illustrate the potential of SA-GCN.

4. Experiments

4.1. Datasets

NTU-60 RGB+D Dataset. This is currently the largest and most widely used indoor action recognition dataset, which contains 56,000 action clips across 60 action categories [29]. The clips were performed by 40 volunteers in various age groups ranging from 10 to 35 years old. Each action was obtained by three cameras at the same height with different horizontal angles: $-45°$, $0°$, and $45°$. This dataset includes location information of 3D joints for each frame detected by Kinect depth sensors, where 25 joints are contained for each subject of the skeleton sequences and each video has no more

than 2 subjects. We evaluate the proposed model using two benchmarks derived from the NTU-60 RGB+D dataset, according to the metrics introduced in [25]. The two benchmarks are:

1) Cross-subject (X-Sub/CS): The dataset based on this benchmark is divided into a training set including 40,320 videos and a validation set containing 16,560 videos, where the persons within the videos are different.

2) Cross-view (X-View/CV): The dataset based on this benchmark has the cameras with different horizontal angles, where 37,920 videos in the training set are obtained from the angles (0° and 45°), and 18,960 videos in the validation set are captured from the angle −45°) .

NTU-120 RGB+D Dataset. This is a large-scale dataset for 3D skeleton-based action recognition, which is obtained from 106 distinct subjects and contains more than 114 thousand video samples and 8 million frames [30]. In fact, this dataset can be seen as an expansion of the NTU-60 RGB+D dataset in the number of performers and action categories. It has 120 different action classes including daily actions, mutual behaviors, and health-related activities. Two specific benchmarks are derived from this dataset, namely X-Sub and cross-setup.

1) X-Sub: The dataset consists of training and testing groups, where each group contains 53 subjects.

2) X-View: The dataset divides samples into training and testing groups, in which the even setup samples are used for training and the odd setup samples are used for testing.

We follow this rule and give the top-1 accuracy about X-Sub and X-View in all experiments.

4.2. Training details

The model proposed in this paper is composed of nine SA-GCN stacks. The first 4 SA-GCN channels of the model are 64, the number of channels is doubled from the 5th to the 7th, and the last three are 256. The convolution stride is set to 2. For fairness in the experimental comparisons, the parameters and number of channels used in different comparative models are the same. Specifically, we train the model for 80 epochs using stochastic gradient descent (SGD) with Nesterov momentum (0.9), batch size 24, and initial learning rate 0.1. In the tenth and fifth iterations, the learning rate is reduced tenfold, respectively. For video data with more than two people, only the first two people are selected, and all skeleton data is filled to T = 300. In addition, this paper chooses a data preprocessing method similar to ST-GCN, in which only skeleton data is selected as the comparison data and the influence of other data streams is not considered. All experiments were conducted on PyTorch with 4 TITANX GPUs.

4.3. Comparison with the state of theart methods

In this section, we perform our proposed method on the NTU-60 and NTU-120 datasets to compare with 3 previous state of the art approaches, which include ST-GCN [25], two-stream adaptive GCN (2s-AGCN) [27] and GCN-NAS[†] [46]. For fair comparisons, we use the data extraction method suggested by ST-GCN [25] in all experiments, which is raw data without preprocessing such as regularization.

[†]GCN-NAS: learning GCN for skeleton-based human action recognition by neural searching.

4.3.1. Baselines

a **ST-GCN** [25]: It is a generic representation of skeleton sequences for action recognition by extending GCNs to a spatial-temporal graph model.

b **2s-AGCN** [27]: It is the two-stream adaptive GCN for skeleton-based action recognition, which models both the first-order and the second-order information simultaneously to improve the acuracy of recognition.

c **GCN-NAS** [46]: It is the automatically designed GCN for skeleton-based action recognition based on neural architecture search, which uses a sampling-and memory-efficient evolution strategy to find the optimal architecture for recognition.

d **STAR** [47]: It is an action recognition model based on sparse transformer.

e **TSTE** [48]: It is a two-stream transformer encoder network based on spatio-temporal feature and shape transformation.

f **TAG** [49]: It is a generalization of ST-GCN based on weak feature extraction.

The proposed model improves the spatial convolution in ST-GCN by adding a SA mechanism, which realizes the application of the attention module in GCN and strengthens the ability of the model to extract joint weights.

4.3.2. Experiment results

Table 1 summarizes the results on the two benchmarks of the NTU-60 dataset. The comparison is based on the same data processing method and the same skeleton data flow. We can find that SA-GCN gives the best results on X-view and X-Sub. For instance, compared to ST-GCN [25], the accuracy of SA-GCN is increased by about 3.2% on both X-View and X-Sub. On X-View, the accuracy of SA-GCN is 1.7% higher than that of 2s-AGCN [27] and GCN-NAS [46]. On X-Sub, SA-GCN achieves an improvement of 5.6% and 2.0% than 2s-AGCN [27] and GCN-NAS [46] respectively.

Table 1. Skeleton-based action recognition performance on NTU-60 dataset. We report the accuracy on both X-Sub and X-View benchmarks.

Methods	X-View	X-Sub
ST-GCN [25]	88.3	81.5
2s-AGCN [27]	89.8	79.1
GCN-NAS [46]	89.8	82.7
STAR [47]	89.0	83.4
TSTE [48]	85.3	80.5
TAG [49]	90.0	82.1
SA-GCN(ours)	**91.5**	**84.7**

The results on the NTU-120 are shown in Table 2. SA-GCN is able to outperform ST-GCN and 2s-AGCN on both X-View and X-Sub. Compared with GCN-NAS, the accuracy of the SA-GCN achieves

suboptimal. This is due to the higher complexity of the GCN-NAS model. However, our model does not involve complex parameters and is basically the same as ST-GCN. Meanwhile, it should be emphasized that the attention mechanisms advocated in SA-GCN all share parameters.

In this part, we give the influences of the SA-GCN on NTU60 dataset, which is shown in Table 3. As can be seen from these experiments, the proposed model with SA-GCN generates unanimously more promising results than the other methods.

Table 2. Skeleton-based action recognition performance on NTU-120 dataset. We report the accuracy on both X-Sub and X-View benchmarks.

Methods	X-View	X-Sub
ST-GCN [25]	78.1	75.5
2s-AGCN [27]	77.5	78.0
GCN-NAS [46]	80.5	78.0
STAR [47]	80.2	78.3
TSTE [48]	67.5	66.6
SA-GCN(ours)	79.2	**78.5**

Table 3. Ablation study of the SA module and GCN on NTU60 dataset.

SAM	GCN	X-View	X-Sub
✓	×	80.0	73.0
×	✓	88.3	81.5
✓	✓	91.5	84.7

5. Conclusions and future work

In this work, we have proposed a unified spatio-temporal SA-GCN for low-quality motion video data with fixed viewing angle, where the designed SA module can be regarded as a GCN. Based on this module, the proposed model can extract features efficiently by learning weight between joint points of different scales. Furthermore, the designed SA mechanism not only characterizes the multi-scale dependencies of joints, but also integrates the structural features of the graph and the ability of self-learning fusion features. Moreover, since the parameters in the attention mechanism are shared, the total number of parameters of the model does not increase significantly. Extensive experiments on NTU-60 RGB+D and NTU-120 RGB+D datasets show that the proposed model achieves substantial improvements over mainstream methods. In the future, we will focus on optimization of the model structure and application in random motion videos.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgment

This work was supported in part by the National Nature Science Foundation of China (62072312, 62071303, 61972264), in part by Shenzhen Basis Research Project (JCYJ20210324094009026, JCYJ20200109105832261), the Project of Educational Commission of Guangdong Province (2023KTSCX116) and National Nature Science Foundation of Guangdong Province (2023A1515011394).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. M. Vrigkas, C. Nikou, I. A. Kakadiaris, A review of human activity recognition methods, *Front. Rob. AI*, **2** (2015), 28. https://doi.org/10.3389/frobt.2015.00028
2. Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, J. Liu, Human action recognition from various data modalities: A review, *IEEE Trans. Pattern Anal. Mach. Intell.*, **45** (2022), 3200–3225. https://doi.org/10.1109/TPAMI.2022.3183112
3. W. Lin, M. T. Sun, R. Poovandran, Human activity recognition for video surveillance, in *2008 IEEE International Symposium on Circuits and Systems (ISCAS)*, IEEE, (2008), 2737–2740. https://doi.org/10.1109/ISCAS.2008.4542023
4. W. Hu, D. Xie, Z. Fu, W. Zeng, S. Maybank, Semantic-based surveillance video retrieval, *IEEE Trans. Image Process.*, **16** (2007), 1168–1181. https://doi.org/10.1109/TIP.2006.891352
5. I. Rodomagoulakis, N. Kardaris, V. Pitsikalis, E. Mavroudi, A. Katsamanis, A. Tsiami, et al., Multimodal human action recognition in assistive human-robot interaction, in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, (2016), 2702–2706. https://doi.org/10.1109/ICASSP.2016.7472168
6. K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, **27** (2014).
7. J. Zhu, Z. Zhu, W. Zou, End-to-end video-level representation learning for action recognition, in *2018 24th International Conference on Pattern Recognition (ICPR)*, IEEE, (2018), 645–650. https://doi.org/10.1109/ICPR.2018.8545710
8. M. R. Sudha, K. Sriraghav, S. Manisha, S. G. Jacob, S. Manisha, Approaches and applications of virtual reality and gesture recognition: A review, *Int. J. Ambient Comput. Intell.*, **8** (2017), 1–18. https://doi.org/10.4018/IJACI.2017100101
9. J. Zhu, W. Zou, Z. Zhu, Y. Hu, Convolutional relation network for skeleton-based action recognition, *Neurocomputing*, **370** (2019), 109–117. https://doi.org/10.1016/j.neucom.2019.08.043
10. L. Shi, Y. Zhang, J. Cheng, H. Lu, Skeleton-based action recognition with multi-stream adaptive graph convolutional networks, *IEEE Trans. Image Process.*, **29** (2020), 9532–9545. https://doi.org/10.1109/TIP.2020.3028207

11. K. Cheng, Y. Zhang, X. He, J. Cheng, H. Lu, Extremely lightweight skeleton-based action recognition with shiftgcn++, *IEEE Trans. Image Process.*, **30** (2021), 7333–7348. https://doi.org/10.1109/TIP.2021.3104182
12. M. Wang, X. Li, S. Chen, X. Zhang, L. Ma, Y. Zhang, Learning representations by contrastive spatio-temporal clustering for skeleton-based action recognition, *IEEE Trans. Multimedia*, **26** (2023), 3207–3220. https://doi.org/10.1109/TMM.2023.3307933
13. C. Pang, X. Gao, Z. Chen, L. Lyu, Self-adaptive graph with nonlocal attention network for skeleton-based action recognition, *IEEE Trans. Neural Networks Learn. Syst.*, **2023** (2023), 1–13. https://doi.org/10.1109/TNNLS.2023.3298950
14. M. Trascau, M. Nan, A. M. Florea, Spatio-temporal features in action recognition using 3D skeletal joints, *Sensors*, **19** (2019), 1–15. https://doi.org/10.3390/s19020423
15. P. Geng, X. Lu, C. Hu, H. Liu, L. Lyu, Focusing fine-grained action by self-attention-enhanced graph neural networks with contrastive learning, *IEEE Trans. Circuits Syst. Video Technol.*, **33** (2023), 4754–4768. https://doi.org/10.1109/TCSVT.2023.3248782
16. T. Xu, W. Takano, Graph stacked hourglass networks for 3d human pose estimation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 16105–16114.
17. B. Doosti, S. Naha, M. Mirbagheri, D. J. Crandall, Hope-net: A graph-based model for hand-object pose estimation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 6608–6617.
18. K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, H. Lu, Skeleton-based action recognition with shift graph convolutional network, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 183–192.
19. M. Li, S. Chen, Y. Zhao, Y. Zhang, Y. Wang, Q. Tian, Dynamic multi-scale graph neural networks for 3D skeleton based human motion prediction, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 214–223.
20. S. Zhang, W. Zhao, Z. Guan, X. Peng, J. Peng, Keypoint-Graph-Driven Learning Framework for Object Pose Estimation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 1065–1073.
21. L. Li, W. Zheng, Z. Zhang, Y. Huang, L. Wang, Skeleton-based relational modeling for action recognition, preprint, arXiv:1805.02556, 2018.
22. W. Zheng, L. Li, Z. Zhang, Y. Huang, L. Wang, Relational network for skeleton-based action recognition, in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, (2019), 826–831. https://doi.org/10.1109/ICME.2019.00147
23. Q. Ke, M. Bennamoun, S. An, F. Sohel, F. Boussaid, A new representation of skeleton sequences for 3D action recognition, in *roceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 3288–3297.
24. T. S. Kim, A. Reiter, Interpretable 3D human action analysis with temporal convolutional networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, IEEE, (2017), 20–28.

25. S. Yan, Y. J. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in *Thirty-Second AAAI Conference on Artificial Intelligence*, **32** (2018). https://doi.org/10.1609/aaai.v32i1.12328

26. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, (2017), 30.

27. L. Shi, Y. Zhang, J. Cheng, H. Lu, Two-stream adaptive graph convolutional networks for skeleton-based action recognition, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 12026–12035.

28. C. Wang, C. Deng, On the global self-attention mechanism for graph convolutional networks, in *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, (2021), 8531–8538. https://doi.org/10.1109/ICPR48806.2021.9412456

29. A. Shahroudy, J. Liu, T. Ng, G. Wang, NTU RGB+D: A large scale dataset for 3D human activity analysis, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 1010–1019.

30. J. Liu, A. Shahroudy, M. Perez, G. Wang, L. Duan, A. C. Kot, NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding, in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **42** (2019), 2684–2701. https://doi.org/10.1109/TPAMI.2019.2916873

31. M. Defferrard, X. Bresson, P. Vandergheynst, Convolutional neural networks on graphs with fast localized spectral filtering, in *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, (2016), 29.

32. M. Niepert, M. Ahmed, K. Kutzkov, Learning convolutional neural networks for graphs, in *Proceedings of The 33rd International Conference on Machine Learning*, PMLR, (2016), 2014–2023.

33. B. Li, X. Li, Z. Zhang, F. Wu, Spatio-temporal graph routing for skeleton-based action recognition, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **33** (2019), 8561–8568. https://doi.org/10.1609/aaai.v33i01.33018561

34. T. Li, R. Zhang, Q. Li, Multi scale temporal graph networks for skeleton-based action recognition, preprint, arXiv:2012.02970, 2020. https://doi.org/10.48550/arXiv.2012.02970

35. H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, H. Jégou, Going deeper with image transformers, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), 32–42.

36. H. Zhang, I. Goodfellow, D. Metaxas, A. Odena, Self-attention generative adversarial networks, in *Proceedings of the 36th International Conference on Machine Learning*, PMLR, (2019), 7354–7363.

37. Y. Rao, J. Lu, J. Zhou, Attention-aware deep reinforcement learning for video face recognition, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (2017), 3931–3940.

38. H. Larochelle, G. E. Hinton, Learning to combine foveal glimpses with a third-order Boltzmann machine, in *Advances in Neural Information Processing Systems 23 (NIPS 2010)*, (2010), 23.

39. F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, et al., Residual attention network for image classification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 3156–3164.

40. J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018), 7132–7141.

41. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, et al., Show, attend and tell: Neural image caption generation with visual attention, in *Proceedings of the 32nd International Conference on Machine Learning*, PMLR, (2015), 2048–2057.

42. M. E. Hussein, M. Torki, M. A. Gowayyed, M. El-Saban, Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations, in *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

43. J. Liu, A. Shahroudy, D. Xu, G. Wang, Spatio-temporal lstm with trust gates for 3D human action recognition, *European Conference on Computer Vision*, Springer, Cham, (2016), 816–833. https://doi.org/10.1007/978-3-319-46487-9_50

44. M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, Q. Tian, Actional-structural graph convolutional networks for skeleton-based action recognition, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 3595–3603.

45. C. Chen, X. Zhao, J. Wang, D. Li, Y. Guan, J. Hong, Dynamic graph convolutional network for assembly behavior recognition based on attention mechanism and multi-scale feature fusion, *Sci. Rep.*, **12** (2022), 1–13. https://doi.org/10.1038/s41598-022-11206-8

46. W. Peng, X. Hong, H. Chen, G. Zhao, Learning graph convolutional network for skeleton-based human action recognition by neural searching, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **34** (2020), 2669–2676. https://doi.org/10.1609/aaai.v34i03.5652

47. F. Shi, C. Lee, L. Qiu, Y. Zhao, T. Shen, S. Muralidhar, et al., Star: Sparse transformer-based action recognition, preprint, arXiv:2017.07089, 2021. https://doi.org/10.48550/arXiv.2107.07089

48. H. Zhang, H. Geng, G. Yang, Two-stream transformer encoders for skeleton-based action recognition, in *6th International Technical Conference on Advances in Computing, Control and Industrial Engineering (CCIE 2021)*, Springer, **920** (2022), 272–281. https://doi.org/10.1007/978-981-19-3927-3_26

49. Y. Meng, M. Shi, W. Yang, Skeleton action recognition based on tranformer adaptive graph convolution, in *Journal of Physics: Conference Series*, **2170** (2022), 012007. https://doi.org/10.1088/1742-6596/2170/1/012007

50. W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, et al., The kinetics human action video dataset, preprint, arXiv:1705.06950, 2017. https://doi.org/10.48550/arXiv.1705.06950

51. X. Qin, R. Cai, J. Yu, C. He, X. Zhang, An efficient self-attention network for skeleton-based action recognition, *Sci. Rep.*, **12** (2022), 1–10. https://doi.org/10.1038/s41598-022-08157-5

52. T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, preprint, arXiv:1609.02907, 2016. https://doi.org/10.48550/arXiv.1609.02907

53. Z. Chen, S. Li, B. Yang, Q. Li, H. Liu, Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **35** (2021), 1113–1122. https://doi.org/10.1609/aaai.v35i2.16197

54. Z. Liu, H. Zhang, Z. Chen, Z. Wang, W. Ouyang, Disentangling and unifying graph convolutions for skeleton-based action recognition, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 143–152.

AIMS Press